2020

Clear-Sighted Statistics: Module 5: Statistical Measures

Edward Volchok
CUNY Queensborough Community College

Clear-Sighted Statistics: An OER Textbook

Module 5: Statistical Measures

> Exploratory data analysis [descriptive statistics] is detective work—or counting detective work—or graphic detective work….The processes of criminal justice are clearly divided between the search for evidence—in Anglo-Saxon lands the responsibility of the police or other investigative forces—and the evaluation of the evidence's strength—a matter for juries and judges. In data analysis a similar distinction is helpful. Exploratory data analysis is detective in character. Confirmatory data analysis [inferential statistics] is judicial or quasi-judicial in character.[1]
>
> -- John W. Tukey

I. Introduction

To paraphrase John W. Tukey, descriptive statistics—"exploratory data analysis," in his words—is "detective work." It finds clues. But, it does not "solve" the case using these clues; which is to say, it does not confirm facts. Descriptive statistics typically leads to hypotheses that may warrant further investigation, while inferential statistics—"confirmatory data analysis"—seeks to confirm or, more correctly, refute these hypotheses.[2] Without a detailed and careful review of the descriptive statistics, analysts would not know what to investigate when conducting inferential statistics.

This module will introduce the quantitative measures used in descriptive and inferential statistics. Many of these measures are essential to inferential statistics. You will be shown how to calculate these measures with paper and pencil, a simple handheld calculator, and with Microsoft Excel.

After completing this module, you will understand:

- **Measures of Central Tendency:** The mode, mean, median, weighted mean, geometric mean, and trimmed mean.
- **Measures of Dispersion:** The range, mean absolute deviation, variance, standard deviation, and the coefficient of variation.

- **Measures of Skewness:** The coefficient of 0 > and kurtosis.

- **Measures of Relative Position:** Percentiles, Quartiles, Interquartile Range, and the Five Number Summary.

- **Estimating the Mean and Standard Deviation** of data grouped in frequency distributions.

- **Microsoft Excel Descriptive Statistics Data Analysis Tool.**

You should download the following files that accompany this module:

- 05_BoxAndWhisker.xlsx
- 05_CV_BigMac_Data.xlsx
- 05_DescriptiveStat-ToolPak
- 05_EstimatingMeanSD.xlsx
- 05_Exercises.xlsx
- 05_GeometricMean.xlsx
- 05_MAD.xlsx
- 05_M&M_Colors.xlsx
- 05_Mean_Outlier.xlsx
- 06_Median_Outlier.xlsx
- 05_Mode.xlsx
- 05_Percentiles_Quartiles.xlsx
- 05_Range.xlsx
- 05_TrimmedMean.x
- 05_VAR_SD.xlsx
- 05_WgtMean_Frappachinos.xls

II. Measures of Central Tendency

Measures of central tendency quantify the most typical or common value in a distribution. For example, what is the most common or typical number of children in a family, the most common educational level, or the most common number of credit cards a person has in his or her wallet. Six measures of central tendency will be covered: A) mode, B) mean, C) median, D) weighted mean, E) geometric mean, and F) trimmed mean.

A) Mode

The mode is the most frequently occurring value in a distribution. A big advantage of the mode is that it is the only measure of central tendency that can be used with nominal data. The mode can also be calculated with ordinal, interval, and ratio data. The mode is also not distorted by extreme values like the mean, the most commonly used measure of central tendency. The mode need not be in the "center" of the data. Because a distribution can have no mode, one mode, or more than one mode, it is considered less useful than the median or the mean.

Here is a frequency table for the distribution of colors in a bag of M&Ms candy:

Table 1: Mode of Nominal Data

Color	f
Blue	5
Brown	6
Green	7
Orange	16
Red	13
Yellow	10
Total	57

The mode, or modal value, for this distribution is **orange**, given that this color appears more frequently than the five other colors. But this conclusion is based on a tiny sample: one bag. We cannot make definitive conclusions about the distribution of colors for this product. We can, however, create hypotheses about this distribution. The data for Table 1 is in 05_M&M_Colors.xlsx.

There is no mathematical formula for calculating the mode. To do this by hand, sort the data in order of magnitude; which means ordered from smallest to highest. When using nominal data, sort the variables alphabetically. Count how many times each value occurs. The mode is the value that appears most often. See the example in Table 2.

Table 2: Mode for Ratio Data

X	X Sorted
1	**1**
3	***2***
4	***2***
2	***2***
9	**3**
2	**4**
6	**6**
9	***9***
2	***9***

This distribution, or set of observations, has only one mode, 2, because it appears three times; 9 is not the mode because it appears only twice.

A distribution can have no mode, one mode, or two or more modes as shown in Table 3.

Table 3: Three Distributions – No Mode, One Mode, and 2 Modes

No Mode	1 Mode	2 Modes
1	1	1
2	***2***	***2***
3	***2***	***2***
4	***2***	***2***
5	3	3
7	4	4
8	6	***9***
9	9	***9***
10	9	***9***

While there is no mathematical formula for calculating the mode, Microsoft Excel has three mode functions:

1. MODE
2. MODE.SNGL
3. MODE.MULT

The syntax for the MODE function is =MODE(number1,[number2]…). Number2 is optional. The arguments can be numbers or cell references that contain numbers. When the data do not contain a mode, Excel returns the #N/A error message.

The MODE function, which is Excel's original mode function, has a major shortcoming. It can only report one mode even when the distribution is multi-modal. With Excel 2010, Microsoft introduced the MODE.SNGL. This function operates just like the MODE function, and has the same shortcoming: It only reports one mode for multi-modal distributions.

The MODE.MULT function, which was also introduced with Excel 2010, overcomes the shortcoming of the MODE and MODE.SNGL functions. It can find multiple modes. The syntax for the MODE.MULT function is = MODE.MULT(number1,[number2]…). The MODE.MULT function must be entered as an *array function*. To enter an array function, highlight several cells where the results are to be placed, type the formula in the first cell and then press the CONTROL, SHIFT, and ENTER (or RETURN) keys at the same time. In Figure 1, fifty variables are shown in Cells A1:E10. The MODE.MULT function is entered in five cells, G4 through G8 as an array. The curly brackets,{}, indicate that the formula is an array. Excel found all three modes**: 143**, **132**, and **142** in this distribution. The #N/A error appears in two cells G7 and G8, which signifies that there is no fourth or fifth mode. The MODE function is entered in G2 and the MODE.SNGL function is in G3. Both functions return only one mode, 143.

	A	B	C	D	E	F	G	H
1	$170	$143	$178	$184	$132	**Generic Formulas**	**Result**	**Formula Used**
2	$177	$141	$145	$164	$125	=Mode(Number1,Number2)	143	=MODE(A1:E10)
3	$129	$123	$159	$175	$132	=MODE.SNGL(Number1,Number2)	143	=MODE.SNGL(A1:E10)
4	$106	$120	$160	$162	$183	=MODE.MULT(Number1,Number2)	143	{=MODE.MULT(A1:E10)}
5	$150	$165	$180	$147	$139		132	{=MODE.MULT(A1:E10)}
6	$124	$148	$137	$146	$149		144	{=MODE.MULT(A1:E10)}
7	$143	$144	$142	$138	$140		#N/A	{=MODE.MULT(A1:E10)}
8	$119	$133	$135	$153	$155		#N/A	{=MODE.MULT(A1:E10)}
9	$179	$158	$130	$157	$113			
10	$111	$144	$154	$166	$151			

Figure 1:The Results Derived from the Three Mode Functions

The data for Figure 1 can be found in 05_Mode.xlsx.

B) Mean

The mean, or arithmetic mean, is the most commonly used measure of central tendency. Many people call this measure the "average," **Please note:** All measures of central tendency are types of averages.

The mean is calculated by adding all of the variables in the data and then dividing the sum of the variables by the number of variables. Equation 1 shows the general formula for calculating the mean:

$$\text{Mean} = \frac{\text{Sum of the Variables}}{\text{Number of Variables}}$$

Equation 1: General Formula for the Mean

When using symbols to write the formula for the mean there are, in fact, two formulas. One for the population mean (μ or the lower case Greek letter "mu") and the other for the sample mean ($\bar{X}$, which is pronounced "X-Bar"). Table 4 shows these formulas:

Table 4: Formulas for the Population Mean and Sample Mean

Population Mean	**Sample Mean**
$\mu = \frac{\Sigma(X)}{N}$	$\bar{X} = \frac{\Sigma(X)}{n}$
Where: μ: Population mean **Σ: Operation of addition** **X: Variables in the population** **N: Number of values in the population**	**Where: $\bar{X}$: Sample mean** **Σ: Operation of addition** **X: Variables in the sample** **n: Number of values in the sample**

Please note: Σ is the capital Greek letter "sigma." This symbol appears in many statistical formulas. It stands for the process of summation or addition.

The formula for the mean in Microsoft Excel is:

=AVERAGE(number1,[number2],…)

Equation 2: Excel's AVERAGE Function

With inferential statistics, we shall see that there is a difference between the sample mean, $\bar{X}$, and the population mean, μ. This difference is called random sampling error.

Calculating the Mean:

Here is a small set of variables. It does not matter whether the variables are considered a population or a sample because μ and $\bar{X}$ would have the same value. A person owns four small dogs. Here, in Figure 2, are the weights of the dogs, and how the mean is calculated:

	A	B	C	D
1	**Pet's Name**	**Animal**	**Weight**	**Formula**
2	John	Maltese	6	
3	Paul	Yorkie	5	
4	Ringo	Chihuahua	4	
5	George	Toy Poodle	7	
6		**Σ**	22	=SUM(C2:C5)
7		**n**	4	=COUNT(C2:C5)
8		**Mean**	5.5	=AVERAGE(C2:C5)

Figure 2: Calculating the Mean

The mean weight for the four dogs, 5.5 pounds. It is the sum of the weights of the four dogs, 22 pounds, divided by the number of dogs, 4. The calculation in Excel is shown in cell D8, =AVERAGE(C2:C5). You could also find the mean by dividing 22 by 4, or =C6/C7. This data can be found in 05_Mean_Outlier.xlsx under the Mean tab.

The Four Characteristics of the Mean

1. The **mean requires quantitative data**; that is to say, interval and ratio data. You cannot calculate the mean for qualitative (nominal and ordinal) data.

2. Every distribution of interval and ratio data has **only one mean**.

3. The **sum of the deviations from the mean always equals zero**.

4. **Outliers distort the mean.** Outliers are variables that are significantly larger or smaller than most other variables in a distribution.

Characteristic 1: The mean cannot be calculated from ordinal data

The frequency table shown in Table 5 shows the order of finish for the New York Yankees from 1903 to 2018.

Table 5: New York Yankees Order of Finish – 1903 to 2018

Place	f
1st Place	47
2nd Place	23
3rd Place	13
4th Place	11
5th Place	10
6th Place	4
7th Place	4
8th Place	2
9th Place	1
10th Place	1
Total	116

Source: http://mlb.mlb.com/nyy/history/year_by_year_results.jsp

You cannot create an average of ordinal data: 1st, 2nd, 3rd, and so on even when you have the raw data.

Characteristic 2: Every distribution of interval and ratio data has only one mean

The formula for the mean, $\Sigma(X)/n$, allows for only one mean. All distributions of interval and ratio data must have one mean.

Characteristic 3: The Sum of the Deviations from the Mean Equals Zero

The third characteristic of the mean is that the sum of the deviations from the mean always equals zero, $\Sigma(X - \text{Mean}) = 0$. The illustration in Table 6 demonstrates this characteristic.

Table 6: Sum of the Deviations from the Mean Equals Zero

X	X - Mean
8	8 - 5 = 3
4	4 – 5 = -1

	3	3 – 5 = -2
Σ	15	0
N	3	
Mean	5	

The mean of the three variables equals 5, found by:

$$(8 + 4 + 3)/3 = 15/3 = 5$$

When the mean is subtracted from each of the three variables, we get 3, -1, and -2. The sum of these three numbers is zero.

This property of the mean raises an issue that will be addressed when the measures of dispersion based on the mean are presented. These measures are the mean absolute deviation, variance, and standard deviation and are used to measure how the data varies from the mean.

Characteristic 4: Distortion of the Mean by Outliers

The mean's susceptibility to distortion by *outliers*—very small or very large variables that are different than most of the data—can be a major disadvantage. Return to our example of the four little dogs. The owner acquires a fifth dog. This one is not a lap dog. It is a 180-pound Saint Bernard named Yoko. The total weight of the five dogs is now 202 pounds, and the mean jumps from 5.5 pounds to 40.4 pounds. Yoko, the 180-pound Saint Bernard, distorts the mean. If Yoko were a 2,200-pound horse, the mean would be 444.4 pounds, and if she were a 10,000-pound elephant, the mean would be 2,000.4 pounds. This data can be found in 05_Mean_Outlier.xlsx under the Mean tab.

	A	B	C	D	E	F	G	H	I	J
1	**Pet's Name**	**Animal**	**Weight**	**Pet's Name**	**Animal**	**Weight**	**Animal**	**Weight**	**Animal**	**Weight**
2	John	Maltese	6	John	Maltese	6	Maltese	6	Maltese	6
3	Paul	Yorkie	5	Paul	Yorkie	5	Yorkie	5	Yorkie	5
4	Ringo	Chihuahua	4	Ringo	Chihuahua	4	Chihuahua	4	Chihuahua	4
5	George	Toy Poodle	7	George	Toy Poodle	7	Toy Poodle	7	Toy Poodle	7
6		**Σ**	22	Yoko	St. Bernard	180	Horse	2,200	Elephant	10,000
7		**n**	4		**Σ**	202	**Σ**	2,222	**Σ**	10,022
8		**Mean**	5.5		**n**	5	**n**	5	**n**	5
9					**Mean**	**40.4**	**Mean**	**444.4**	**Mean**	**2,004.4**

Figure 3: Distortion of the Mean By Outliers

While the mean is the most widely used measure of central tendency, another measure of central location is used in certain cases because of the mean's susceptibility to distortion by outliers. When the inclusion of outliers cannot be avoided, the median is used. We will now turn to this measure.

C) Median

The median is the middle value *when the data are ranked from smallest to largest*. There are as many values above the median as below it. Because the median is based on the value or values in the middle of the distribution and not on all values, it is not distorted by outliers. Measures that are not affected by outliers are called *robust*. The median is sometimes symbolized as M, Md, or $\tilde{x}$ (x-tilde). The median can be calculated with ordinal, interval, or ratio data. Unlike the mode and like the mean, a distribution can have only one median.

Calculating the Median

John Tukey presented this "formula" for the median:

Table 7: John Tukey's Formula for the Median[3]

$$\text{Median} = \begin{cases} \text{The single middle value (when n is an odd number)} \\ \text{The mean of the two middle values (when n is an even number)} \end{cases}$$

When the number of variables in the data is odd, the median is the middle value of the ranked data. When the number of variables is even, the median is the mean of the two values that surround the middle of the ranked data.

Here is how to calculate the median for an odd set of values. In our example there are only five numbers: 10, 4, 15, 2,000, and 2.

Step 1: Sort the data in ascending order: 2, 4, 10, 15, 2,000.

Step 2: Find the number in the middle of the sorted distribution. Given that there are five numbers, the middle value would be the third value from the bottom of the distribution or the third value from the top of the distribution: 2, 4, **10**, 15, 2,000. **The median is 10.**

Finding the median for an even set of values is slightly different.

Step 1: Sort the data in ascending order: 2, 4, 10, 15, 21, 2,000.

Step 2: Find the two values surrounding the middle using the following formula, (n + 1)/2. With six variables the formula would be (6 + 1)/2 = 7/2 = 3.5. The median is half way between the third and fourth variables: 2, 4, **10, 15**, 21, 2,000.

Step 3: The third and fourth variables are 10 and 15. To find the half-way point, take the mean of 10 and 15, [10 + 15) = 25/2 = 12.5. **The median is 12.5.**

Calculating Median in Excel

Calculating the median in Excel is very easy using the MEDIAN function. **With the MEDIAN function there is no need to sort the data.** This is a big advantage. The syntax for this function is =MEDIAN(number1,[number2],…). The arguments for the median function can be numbers or cell references that contain numbers. **Remember:** If you are calculating the median by hand, the data ***must*** be sorted in ascending or descending order.

To see how the median is not distorted by outliers like the mean, look at Figure 4 below. The medians are in cells C9, F9, H9, and J9.

	A	B	C	D	E	F	G	H	I	J
1	**Pet's Name**	**Animal**	**Weight**	**Pet's Name**	**Animal**	**Weight**	**Animal**	**Weight**	**Animal**	**Weight**
2	John	Maltese	6	John	Maltese	6	Maltese	6	Maltese	6
3	Paul	Yorkie	5	Paul	Yorkie	5	Yorkie	5	Yorkie	5
4	Ringo	Chihuahua	4	Ringo	Chihuahua	4	Chihuahua	4	Chihuahua	4
5	George	Toy Poodle	7	George	Toy Poodle	7	Toy Poodle	7	Toy Poodle	7
6		**Σ**	22	Yoko	St. Bernard	180	Horse	2,200	Elephant	10,000
7		**n**	4		**Σ**	202	**Σ**	2,222	**Σ**	10,022
8		**Mean**	5.5		**n**	5	**n**	5	**n**	5
9		**Median**	**5.5**		**Mean**	**40.4**	**Mean**	**444.4**	**Mean**	**2,004.4**
10			=MEDIAN(C2:C5)		**Median**	**6.0**		**6.0**		**6.0**
11						=MEDIAN(F2:F6)		=MEDIAN(H2:H6)		
12									=MEDIAN(HJ:J6)	

Figure 4: The Median is not distorted by outliers

You will notice that the median for the four small dogs is the same as the mean, 5.5 pounds. But when we add Yoko, who is either a Saint Bernard, a horse, or an elephant, the median is only 6 pounds. This is because the median is calculated on the middle value or values. Yoko could even be a planet the size of Jupiter, which weighs approximately 4.184 x 10^27 pounds (that is 4.184 followed by 24 zeros), and the median for the four dogs and the planet Yoko would still be 6 pounds. The changing size of Yoko in Figure 4 shows that **unlike the mean, outliers do not distort the median.** This property is why the median should always be considered when developing descriptive statistics. The median is the most commonly used measure of central tendency when data are prone to distortion by outliers. This includes the averages for selling price of homes, people's annual incomes, or their net worth.

D) The Positions of the Mean, Median, and Mode

The positions of the mean, median, and mode depend on the shape of the distribution.

Symmetrical Distributions:

In a **symmetrical distribution** like those shown in Figure 5, the mean, median, and mode are equal and are located at the center of the distribution.

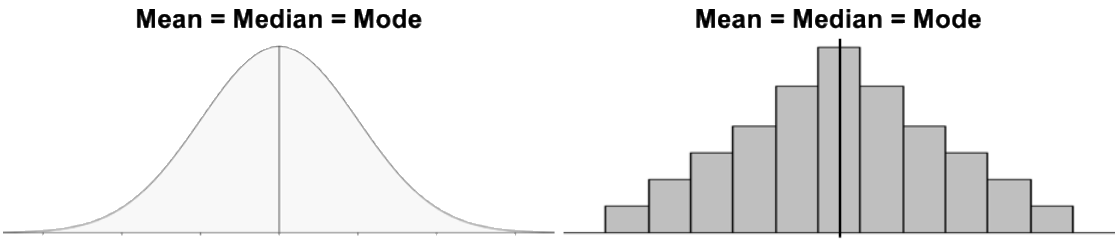

Figure 5: Placement of the mean, median, and mode in continuous and discrete symmetrical distributions

Right Skewed Distributions:

When a distribution is not symmetrical around the mean, it is skewed. In a **right or positive skewed distribution** like the ones shown in Figure 6, the Mode is at the peak of the curve, the median is at the geographic center of the curve, and the mean is pushed to the right by outliers in the right tail.

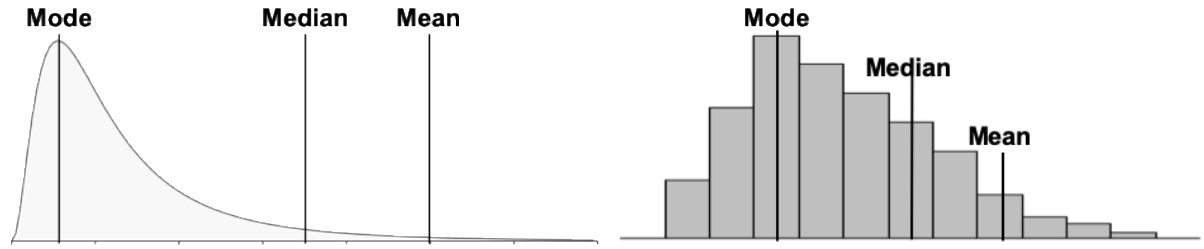

Figure 6: Placement of the mode, median, and mean in continuous and discrete right skewed distributions

Left Skewed Distributions:

In a **left or negative skewed distribution** like the ones shown in Figure 7, the Mode is at the peak of the curve, the median is at the geographic center of the curve, and the mean is pushed to the left by outliers in the left tail.

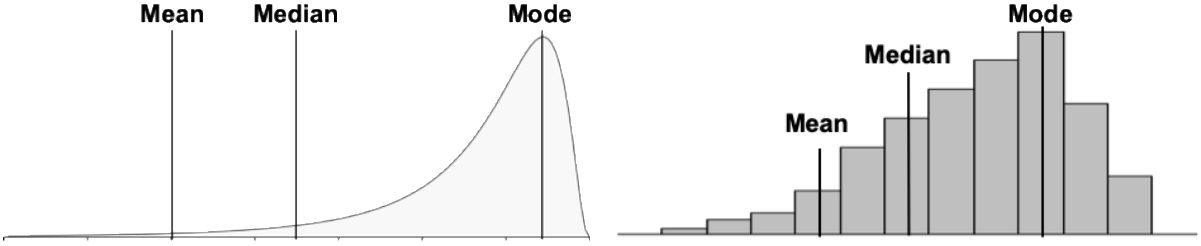

Figure 7: Placement of the mean, median, and mode in continuous and discrete left skewed distributions

Coefficient of Skewness

Skewness can be measured using the coefficient of skewness, which is also known as Pearson's coefficient of skewness after the statistician Karl Pearson. The formula for the coefficient of skewness is:

Equation 3: Formula for the Coefficient of Skewness

$$\text{Coefficient of Skewness} = \frac{3(\bar{X} - \text{Md})}{s}$$

Where: $\bar{X}$: Sample Mean
Md: Median
s: Sample standard deviation

The coefficient of skewness can use either the mode or the median. A word of caution is warranted about using the mode. The coefficient of skewness will not be a stable measure of central location when the data is multimodal or has no mode. Using the median, therefore, is preferred.

Here are the basic guidelines for interpreting the coefficient of skewness. The direction of the skew is given by the sign. Negative values indicate a left or negative skew. Positive values indicate a right or positive skew. A value at or new zero indicates little or no skew.

Table 8: Interpreting the Coefficient of Skewness

Interpretation	Left Skew	Right Skew
Very Good Symmetry	0 > -0.10	0 < 0.10
Good Symmetry	-0.10 > -0.20	0.10 < 0.20
Acceptable Symmetry	-0.20 > -0.30	0.20 < 0.30
Poor Symmetry	<-0.30	>0.30

Table 9 shows players' batting averages for the 2018 American League Championship Series between the New York Yankees and the Boston Red Sox.

Table 9: 2018 ALCS Batting Average

New York Yankees	Boston Red Sox
0.111	0.286
0.000	0.188

0.214	0.294
0.000	0.167
0.200	0.286
0.375	0.667
0.133	0.308
0.200	0.000
0.200	0.357
0.308	0.333
0.231	0.182
0.250	0.333
-	0.000
-	0.333
S = 0.111	S = 0.164
$\bar{X}$ =0.185	$\bar{X}$ = 0.267
CS = -0.401	CS = -0.426

Based on the coefficients of skewness, the distribution of batting averages for both teams are skewed. The New York Yankees have a strong negative skew, -0.401. The Boston Red Sox have a slightly stronger negative skew, -0.426. This means the data is distorted to the left by players with low batting averages.

Please note: Microsoft Excel's SKEW function does not calculate Pearson's coefficient of skewness. Excel's SKEW function uses the third power of deviations around the mean. The syntax for this function is =SKEW(Number1,number2,…). The arguments can be numbers of cell references that contain quantitative data.

If you are not a baseball fan, you may ask, **what is a batting average?** This is, in fact, a statistics question. A batting average (BA) is defined as the number of "hits" divided by the number of "at bats." At bats are when a batter reaches base via a fielder's choice, hit, an error (not including catcher's interference). or when a batter is put out on a non-sacrifice. Batting averages are reported as a decimal with three places. A batting average of .300, three hits out of ten at bats, is considered excellent. The lowest possible batting average is

.000, which means that the batter failed to get a hit at any at bats. The highest possible batting average is 1.000, which would mean the batter always gets a hit.

The batting average was invented by Henry Chadwick, a nineteenth century writer and statistician. He was admitted to the Baseball Hall of Fame in 1938.[4]

Kurtosis

Kurtosis is another measure developed by the great statistician Karl Pearson. It measures the extent to which a distribution differs from a normal or bell-shaped curve by measuring the thickness of the curve's tails.[5] It does not, as many people think, measure the "peakedness," or height, of a curve. Kurtosis is expressed graphically and numerically. Kurtosis can be positive or negative. A negative number means that the curve has thinner tails than a normal curve, a positive number means the curve has fatter tails than a normal curve.

There are three types of kurtosis.

1. **Mesokurtic**

   Mesokurtic distributions are normal distributions or nearly normal distributions. Mesokurtic distributions have zero or close to zero kurtosis.

2. **Platykurtic**

   Platykurtic distributions have negative kurtosis. These distributions have thin tails, which indicates a small number of outliers in a distribution. Investors like platykurtic distributions because the returns on their investments are more predictable.

3. **Leptokurtic**

Leptokurtic distributions have positive kurtosis. These distributions have fat tails, which indicates a larger number of outliers than a mesokurtic distribution.

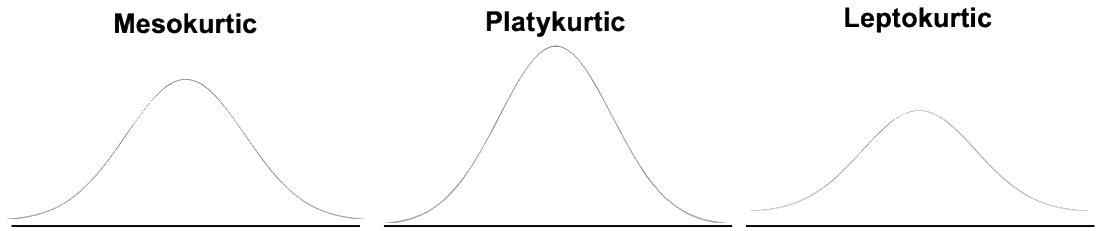

Figure 8: Mesokurtic, Platykurtic, and Leptokurtic Distributions

Excel has a kurtosis function, KURT. The syntax for this function is =KURT(Number1,number2,…). The arguments can be numbers or cell references that contain numbers. Because kurtosis is not often used in elementary statistics, the formula for calculating sample or population kurtosis by hand will not be shown.

E) Weighted Mean

With the arithmetic mean, every variable has equal weight. What happens when variables have unequal weights? The arithmetic mean cannot be used. The weighted mean, not the arithmetic mean, must be used. The formula for the weighted mean is:

Equation 4: Weighted Mean Equation

$$\bar{X}_w = \frac{\Sigma(wX)}{\Sigma w}$$

Where: $\bar{X}_w$: Weighted mean
X: Random variables
w: Weights
Σ: Operation of addition

Here is an example of how the weighted mean is used. Starbucks sells Frappucchinos in three sizes: Tall (small), grande (medium), and venti (large). Here are the prices for each size: Tall, $3.95; grande, $4.45; and venti, $4.95. Last weekend at the Starbucks on Main Street and 34th Avenue, Starbucks sold 1,000 talls, 1,750 grandes, and

1,400 ventis. **What is the average price per cup sold?** The arithmetic mean cannot be used because the variables have different weights; which is to say, different unit sales.

If Starbucks sold exactly the same number of talls, grandes, and ventis, the average price per cup sold would be found using the arithmetic mean:

$$\Sigma(X)/n = (\$3.95 + \$4.45 + \$4.95)/3 = \$13.35/3 = \$4.45$$

Equation 5: Incorrect Calculation of the Weighted Mean

But, the sales of the three sizes are not equal. The formula for the weighted mean must be used. Once it has been determined that the weighted mean must be used, two questions arise: 1) What are the X values and what are the weights? 2) To determine the X values, simply look at the problem. The question to be answered is: **What was the average price per cup sold?** The X values, therefore, are the prices. **The weights are how many of the X variables are in the data.**

Table 10: Weighted Mean Price for Starbucks Frappucchinos

Item	w	X	wX
Tall	1,000	$3.95	$3,950.00
Grande	1,750	$4.45	$7,787.50
Venti	1,400	$4.95	$6,930.00
Total	4,150		$18,667.50
			4,150

Weighted Mean = $4.50 found by $18,667.50/4,150

The data for Tables 9, 10, and 11 can be found in 05_WgtMean_Frappachinos.xls.

There are two common mistakes made when trying to solve a problem like this. The first mistake is to calculate the arithmetic mean, $4.45. The second mistake is confusing the w and X variables. Table 11 shows the result of the second mistake.

Table 11: The serious mistake of confusing the w and X variables

Item	w	X	wX
Tall	$3.95	1,000	$3,950.00
Grande	$4.45	1,750	$7,787.50
Venti	$4.95	1,400	$6,930.00

Total	$13.95	$\underline{\$18{,}667.50}$	**Weighted Mean = $1,381.31**
		$13.95	Found by $18,667.50/$13.95

The second mistake is easy to catch if you are alert and read the problem carefully. Before we perform any calculations, we know that the average price per cup sold must be above $3.95 and below $4.95, found by looking at the lowest and highest price. The calculated price of $1,381.33 per cup is obviously wrong. Why should the mean cup cost 278 times more than the highest priced cup? Follow this rule when calculating the weighted mean: **If the calculated answer for the weighted mean is outside of the range of the highest or lowest X variables, the answer is wrong.**

F) Geometric Mean

The geometric mean is a measure of central tendency used to find the average change in percentages, ratios, indices, and growth rates over time. Indices will be introduced in Module 6. The geometric mean is commonly used to determine the performance of an investment portfolio.

The geometric mean is technically defined as "the *nth* root product of *n* numbers." Here is the formula for the geometric mean:

Equation 6: Geometric Mean

$$GM = \sqrt[n]{(1 + X_1)(1 + X_2) \dots (1 + X_n)}$$

Because the formula requires *nth* roots, the geometric mean is very difficult to calculate using an inexpensive handheld calculator. Fortunately, calculating the geometric mean is easy using Microsoft Excel's GEOMEAN function.

The syntax for the GEOMEAN function is:

=GEOMEAN(number1, [number2], ...)

Equation 7: Excel's GEOMEAN Function

The argument **for this function is Number1, number2, ….**Number1 is required, subsequent numbers are optional.

The geometric mean has a number of advantages:

- It more accurately determines average changes in variables over time than the arithmetic mean.
- All values are used.
- It is not effected by outliers.
- It is not effected by fluctuations in the sample data.

Please Note:

1. **The geometric mean will always be less than or equal to the arithmetic mean.**
2. **All variables must be non-negative numbers**. To get around this problem, the X values are converted to index numbers.

Here is an example of how the geometric mean gives a better result than the arithmetic mean when calculating the mean rate of return for an investment. An investor makes an investment of $10,000. The first year the investment earns 15 percent, the second year 12 percent, the third year 9 percent, and the fourth year 7 percent. As shown in Figure 9, at the end of Year 1, the investment is worth $11,500. By the end of Year 4 the $10,000 initial investment is worth $15,021.94. **What is the average annual increase in value?**

	A	B	C	D	E	F	G
1	Year	Rate of Return	Starting Value	Return $	Closing Value	% Change	% Change
2	1	15.00%	$10,000.00	$1,500.00	$11,500.00	115.00%	15.00%
3	2	12.00%	$11,500.00	$1,380.00	$12,880.00	112.00%	12.00%
4	3	9.00%	$12,880.00	$1,159.20	$14,039.20	109.00%	9.00%
5	4	7.00%	$14,039.20	$982.74	$15,021.94	107.00%	7.00%
6						**Geometric Mean:**	**10.32%**
7							=GEOMEAN(G2:G5)
8						**Arithmetic Mean:**	**10.75%**
9							=AVERAGE(G2:G5)

Figure 9: Geometric Mean vs. Arithmetic Mean

Using the geometric mean, the average annual increase is 10.32%. This result is more accurate than the arithmetic mean of 10.75%. The data for this calculation is in 05_GeometricMean.xlsx.

G) Trimmed Mean

The trimmed mean is used to remove the impact of outliers on the mean. The trimmed mean, also known as the *truncated* or *adjusted mean*, removes a small percentage of the largest and smallest values before calculating the mean. After a pre-selected percentages of values are removed, the trimmed mean is found using the standard arithmetic mean formula. Trimmed means are often used in economics because it evens out the results and thereby provides a more realistic picture of the data.

Microsoft Excel has a handy trimmed mean function, TRIMMEAN. This function returns the mean after the users sets the percentage of the variables to be excluded from the top and bottom tails of the ranked distribution. The benefit of using this function is you do not have to sort the data.

Here is the syntax for TRIMMEAN: =TRIMMEAN(Array,Percent). Array is the range of values or cells containing quantitative data to be trimmed and Percent is the proportion to be trimmed.

	A	B	C	D
1		**Data**	**Formula**	
2		1		
3		5		
4		62		
5		15		
6		55		
7		69		
8		110		
9		35		
10		25		
11		25		
12		45		
13		50		
14		32		
15	Mean	40.69	=AVERAGE(B2:B14)	
16	Trimmed Mean 10%	40.69	=TRIMMEAN(B2:B14,0.1)	5% trimmed on each end. No values omitted.
17	Trimmed Mean 20%	38.00	=TRIMMEAN(B2:B14,0.2)	10% trimmed on each end.
18	Trimmed Mean 35%	38.22	=TRIMMEAN(B2:B14,0.35)	17.5% trimmed on each end.

Figure 10: Microsoft Excel's Trimmed Mean

The data for the trimmed mean can be found in 05_TrimmedMean.xlsx.

III. Measures of Dispersion

Measures of dispersion typically, but not always, describe how variables are distributed from the "center" of the data. The smaller the measure of dispersion, the less dispersion from the mean or median. The following section covers the following measures of dispersion: *Range*, *mean absolute deviation*, *variance*, and *standard deviation*. The *empirical* or *normal rule*, which is based on the mean and standard deviation of a distribution will be covered, as will the coefficient of variation.

A) Range

The range is the simplest measure of dispersion for quantitative data (interval or ratio data). It does *not* measure how far the data varies from a central location measure. The range is the difference between the largest and smallest values in a distribution. Equation 8 shows the formula for calculating the range:

$$R = H - L$$

Equation 8: Formula for the Range

Where: R: Range
H: Highest value
L: Lowest value

Figure 11 shows U.S. grocery sales from 1992 through 2018. The range is found by subtracting the smallest value, $337.37 billion from the largest value, $641.04 billion: $303.67 billion. Excel will also calculate the range when Excel's MAX and MIN functions are combined. Using this U.S. Census data on grocery store sales, the range is calculated with the following formula: =MAX(B2:B27)-MIN(B2:B27). The data for Figure 11 can be found in 05_Range.xlsx.

	A	B	C	D
1	**Year**	**Grocery Store Sales, Billion $**	**Range**	**Formula**
2	1992	**$337.37**	$303.67	=MAX(B2:B27)-MIN(B2:B27)
3	1993	$341.32		
4	1994	$350.52		
5	1995	$356.41		
6	1996	$365.55		
7	1997	$372.57		
8	1998	$378.19		
9	1999	$394.25		
10	2000	$402.52		
11	2001	$418.13		
12	2002	$419.81		
13	2003	$427.99		
14	2004	$441.14		
15	2005	$457.67		
16	2006	$471.70		
17	2007	$491.36		
18	2008	$511.39		
19	2009	$510.33		
20	2010	$520.75		
21	2011	$547.48		
22	2012	$563.65		
23	2013	$574.55		
24	2014	$599.60		
25	2015	$613.23		
26	2016	$626.98		
27	2017	**$641.04**		
28	Source:	U.S. Census Bureau, April 2018		

Figure 11: U.S. Grocery Sales, 1992-2017

The advantage of the range is that it is very easy to calculate and it provides a clear measure of how spread out the data are. The major weakness of the range as a measure of dispersion stems from its simplicity. **It only uses two variables:** The smallest and largest variables in a distribution. Due to the fact that the range does not use all of the data, it

provides little information on how *all* the data are dispersed throughout the distribution.

Because of this, the range is not considered a particularly useful measure of dispersion. Statisticians rely on measures of dispersion that use all the data. That said, a variant of the range, the *interquartile range*, in combination with other measures, is very informative. The interquartile range will be reviewed shortly.

B) Mean Absolute Deviation

One measure of dispersion that uses all the data is the *mean absolute deviation* or MAD. Sometimes this measure is called the *average deviation*, *mean deviation* or *absolute deviation*. MAD is the "average" distance of each variable from the mean.

The calculation of MAD is based on the absolute value of each variable from the mean. The absolute value of a number treats negative numbers as if they were positive. Doing so overcomes the third property of the mean: **The sum of the deviations from the mean always equals zero**.

The larger the MAD, the more dispersed the data. The smallest value for the MAD would be zero, in which case all values would be equal to the mean. As a measure of dispersion, MAD has two advantages over the range: 1) All of the data are used and 2) It is not unduly influenced by outliers. MAD, however, has serious disadvantages: 1) Negative signs are ignored while calculating deviations from the mean, and 2) it is not capable of further algebraic treatment. Because of this MAD is seldom used. Two closely related measures—variance and standard deviation—are widely used to measure dispersion in populations and samples.

The formula for MAD is:

$$MAD = \frac{\Sigma|X - \bar{X}|}{n}$$

Equation 9: Formula for the Mean Absolute Deviation

Where: X: Random variables
$\bar{X}$: Sample mean
||: Absolute values
n: Number of observation in the sample
Σ: Operation of addition

The MAD function in Excel is AVEDEV. It uses the following syntax:

=AVEDEV(number1, [number2],…)

Equation 10: Excel's AVEDEV Function

The argument number1 is required. It can be any cell references containing quantitative data.

Figure 12 shows the ages of the nine New York Yankees starting the game played on Tuesday, July 23, 2019. It also shows how the MAD is calculated by hand and by using Excel's AVEDEV function. The mean age for the starting lineup was 27.89 years-old with a MAD of 2.12 years. The data shown in Figure 12 can be found in 05_MAD.xlsx.

	A	B	C	D
1	**Player**	**X (Age)**	**X - Mean**	**|X - Mean|**
2	Aaron Hicks	28	0.11	0.11
3	Aaron Judge	27	-0.89	0.89
4	Edwin Encarnacion	36	8.11	8.11
5	Luke Voit	28	0.11	0.11
6	Didi Gregorius	29	1.11	1.11
7	Gary Sanchez	26	-1.89	1.89
8	Gleyber Torres	22	-5.89	5.89
9	Gio Urshela	27	-0.89	0.89
10	Mike Tuchman	28	0.11	0.11
11	**Σ**		Σ|X - Mean|	19.11
12	**n**	9	n	9
13	**Mean**	27.89		
14			MAD (by hand) =	**2.12**
15			=AVEDEV(B2:B10)	**2.12**

Figure 12: MAD of the Age of the NY Yankees starting lineup on 7/23/19
Sources: Starting lineup: https://www.mlb.com/yankees/roster/starting-lineups. Age of players: https://www.baseball-reference.com/teams/NYY/2019-roster.shtml

Here are the five steps to calculate MAD by hand:

Step 1: Find the mean.

Step 2: Find the deviations of each random variable from the mean, $X - \bar{X}$.

Step 3: Find the absolute value of $|X - \bar{X}|$.

Step 4: Sum the absolute values, $\Sigma|X - \bar{X}|$.

Step 5: Divide $\Sigma|X - \bar{X}|$ by the number of observations, n.

C) Variance and Standard Deviation

Variance and standard deviation are closely linked measures of variability. Standard deviation is the gold standard for measuring the variability of interval and ratio data. Variance, while less important in descriptive statistics than the standard deviation, is used to calculate standard deviation and other measures commonly used in inferential statistics.

Variance

There are two variances: One for populations, σ^2, the other for the samples, s^2. Sample variance corrects for random sampling error that occurs when we use sample data.

Population variance is the sum of the squared deviations from the population mean divided by the number of variables in the population. Sample variance is the sum of the squared deviations from the sample mean divided by the number of variables in the sample minus one; n – 1, which is called *degrees of freedom*.

Table 12: Population Variance (σ^2) and Sample Variance (s^2)

Population Variance, σ^2	**Sample Variance, s^2**
$$\sigma^2 = \frac{\Sigma(X - \mu)^2}{N}$$	$$s^2 = \frac{\Sigma(X - \bar{X})^2}{n - 1}$$
Where: Σ: Operation of addition μ: Population mean X: Variables in the population N: Number of values in the population	Where: Σ: Operation of addition $\bar{X}$: Sample mean X: Variables in the sample n: Number of values in the sample

Figure 13 shows the ages of the nine New York Yankees in the starting lineup for the game played on Tuesday, July 23, 2019. It also shows how the population and sample variance are calculated by hand and by using Excel's variance functions. This data can be found in the workbook 05_VAR_SD.xlsx under the worksheet labeled "Variance." Population variance equals 11.88 while sample variance equals 13.36.

	A	B	C	D	E	F
1	**Player**	**X (Age)**	**(X - Mean)**	**(X - Mean)^2**		
2	**Aaron Hicks**	**28**	**0.11**	**0.01**		
3	**Aaron Judge**	**27**	**-0.89**	**0.79**		
4	**Edwin Encarnacion**	**36**	**8.11**	**65.79**	$\sigma^2 = \frac{\sum(X-\mu)^2}{N}$	
5	**Luke Voit**	**28**	**0.11**	**0.01**		
6	**Didi Gregorius**	**29**	**1.11**	**1.23**		
7	**Gary Sanchez**	**26**	**-1.89**	**3.57**		
8	**Gleyber Torres**	**22**	**-5.89**	**34.68**	$s^2 = \frac{\sum(X-\bar{X})^2}{n-1}$	
9	**Gio Urshela**	**27**	**-0.89**	**0.79**		
10	**Mike Tuchman**	**28**	**0.11**	**0.01**		
11	**Σ**		**Σ(X - Mean)^2**	**106.89**		
12	**n**	**9**	**n**	**9**		
13	**Mean**	**27.89**	**n - 1**	**8**		
14	**Population Variance**		**σ^2 by hand**	**11.88**		
15			**σ^2 by Excel**	**11.88**	=VAR.P(B2:B10)	
16						
17	**Sample Variance**		**s^2 by hand**	**13.36**		
18			**s^2 by Excel**	**13.36**	=VAR.S(B2:B10)	

Figure 13: Population and Sample Variance
Sources: Starting lineup: https://www.mlb.com/yankees/roster/starting-lineups. Age of players: https://www.baseball-reference.com/teams/NYY/2019-roster.shtml

Here are the five steps to calculate variance by hand:

Step 1: Find the mean. **Please Note:** The population and sample means will have the same value.

Step 2: Find the deviations of each random variable from the mean (X – Mean).

Step 3: Square the deviations from the mean, $(X - Mean)^2$.

Step 4: Sum the squared deviations from the mean, $\Sigma(X - Mean)^2$.

Step 5: For population variance, σ^2, divide the squared deviations from the mean by the number of observations. For sample variance, s^2, divide the squared deviations from the mean by the number of observations minus one.

To calculate population variance using Excel, type VAR.P(B2:B10). Excel returns 11.88. To calculate sample variance using Excel, type VAR.S(B2:B10). Excel returns 13.36. The larger the variance the more variable the data; which is to say, the more widely it is dispersed from the mean.

The following points should be noted:

1) **Sample variance will always be larger than population variance when the same data is used.** This is because the denominator is n – 1, not N. This adjustment in the formula causes sample variance to be greater than population variance is done to adjust for the uncertainty caused by random sampling error.

2) **Variance is *not* in the same units as the mean.** The units for variance are the squared units of the mean. In our example, 11.88 and 13.86 are "squared" years. **This is a serious disadvantage** when reporting descriptive statistics. Squared years is a mathematical abstraction that sheds little light on practical issues.

3) Because of this disadvantage, variance is not an important measure for describing data.

4) Variance, however, is a very important measure because it is used to calculate other measures. One of these measures is standard deviation.

Standard Deviation

Standard deviation is the square root of variance. The reason why standard deviation is the most widely used measure of dispersion for interval and ratio data is that it is in the same units as the mean and it uses all the data.

Table 13: Formulas for Population and Sample Standard Deviation

Population Standard Deviation, σ	**Sample Standard Deviation, s**
$\sigma = \sqrt{\frac{\Sigma(X-\mu)^2}{N}}$	$s = \sqrt{\frac{\Sigma(X-\overline{X})^2}{n-1}}$
Where: Σ: Operation of addition	Where: Σ: Operation of addition

μ: Population mean X: Variables in the population N: Number of values in the population	$\bar{X}$: Sample mean X: Variables in the sample n: Number of values in the sample

Figure 14 shows the ages of the starting nine New York Yankees for the game played on Tuesday, July 23, 2019. It also shows how to calculate the population and sample standard deviation by hand and by using Excel's standard deviation functions. This data can be found in in the workbook VAR_SD.xlsx under the worksheet labeled "Standard Dev." Population standard deviation equals 3.45 while sample standard deviation equals 3.66. To calculate population standard deviation using Excel, type STDEV.P(B2:B10). Excel returns 3.45. To calculate sample variance using Excel, type STDEV.S(B2:B10). Excel returns 3.66. Like variance, sample standard deviation is larger than population standard deviation. The smallest standard deviation is zero. This would mean that all variables are equal. The larger the standard deviation, the more variable the data.

	A	B	C	D	E	F
1	**Player**	**X (Age)**	**(X - Mean)**	**$(X - Mean)^2$**		
2	**Aaron Hicks**	**28**	**0.11**	**0.01**		
3	**Aaron Judge**	**27**	**-0.89**	**0.79**		
4	**Edwin Encarnacion**	**36**	**8.11**	**65.79**	$\sigma = \sqrt{\frac{\sum(X-\mu)^2}{N}}$	
5	**Luke Voit**	**28**	**0.11**	**0.01**		
6	**Didi Gregorius**	**29**	**1.11**	**1.23**		
7	**Gary Sanchez**	**26**	**-1.89**	**3.57**		
8	**Gleyber Torres**	**22**	**-5.89**	**34.68**	$s = \sqrt{\frac{\sum(X-\bar{X})^2}{n-1}}$	
9	**Gio Urshela**	**27**	**-0.89**	**0.79**		
10	**Mike Tuchman**	**28**	**0.11**	**0.01**		
11	**Σ**		**Σ(X - Mean)^2**	**106.89**		
12	**n**	**9**	**n**	**9**		
13	**Mean**	**27.89**	**n - 1**	**8**		
14		**Population Variance**		**11.88**	=VAR.P(B2:B10)	
15		**Sample Variance**		**13.36**	=VAR.S(B2:B10)	
16						
17	**Pop. Standard Deviation**		**σ by hand**	**3.45**		
18			**σ by Excel**	**3.45**	=STDEV.P(B2:B10)	
19						
20	**Sample Standard Deviation**		**s by hand**	**3.66**		
21			**s by Excel**	**3.66**	=STDEV.S(B2:B10)	

Figure 14: Variance and Standard Deviation
Sources: Starting lineup: https://www.mlb.com/yankees/roster/starting-lineups. Age of players: https://www.baseball-reference.com/teams/NYY/2019-roster.shtml

The data in Figure 14 can be found in the workbook 05_VAR_SD.xlsx under the worksheet labeled "Standard Dev."

Ronald Fisher, the renowned English statistician, famously stated that statistical methods reduce data to a few measures that represent all the variables.[6] The two most important measures are the mean and standard deviation. This will become apparent when we turn to inferential statistics.

D) The Empirical or Normal Rule

The empirical or normal rule shows the importance of the mean and standard deviation in a normal distribution, which is also called the bell-shaped curve. It is a symmetrical distribution that peaks at the center, which is the location of the mean, as well as the median and mode.

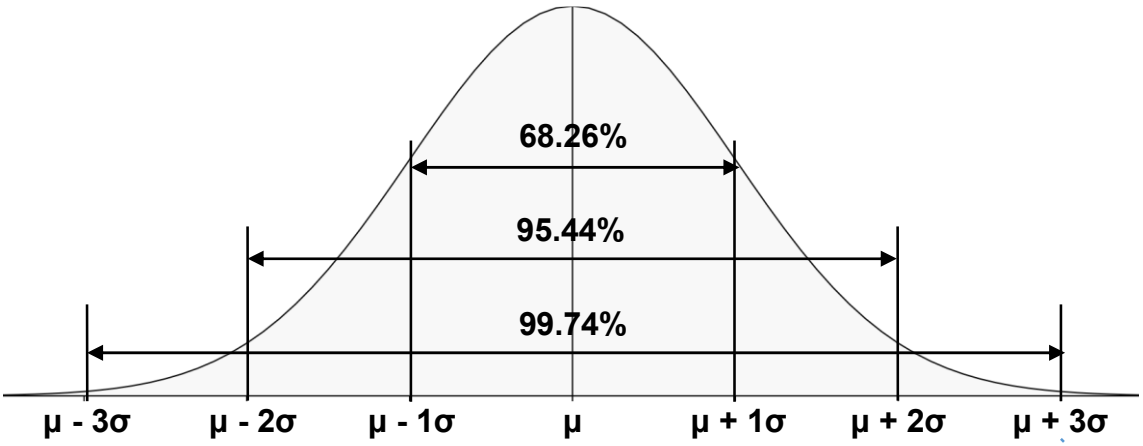

Figure 15: The Empirical or Normal Rule

The empirical rule deals with the distribution of the data in a normal curve. Approximately 68.26 percent of the data are located plus or minus one standard deviation, σ, from the mean, μ, 95.44 percent of the data are located plus or minus two standard deviations from the mean, and 99.74% of the data are located plus or minus three standard deviations from the mean.

Suppose the mean of a distribution is $1,000 and the standard deviation is $100. The empirical rule states:

- 68.26 percent of the variables are between $900 and $1,100 (± 1σ).
- 95.44 percent of the variables are between $800 and $1,200 (± 2σ).
- 99.74 percent of the variables are between $700 and $1,300 (± 3σ)
- 0.13 percent of the variables are above $1,300 (> 3σ).
- 0.13 percent of the variables are below $700 (< 3σ).

The empirical rule will be covered in greater detail in Module 9, Continuous Probability Distributions, and Module 10, Sampling and Sampling Errors. To discuss the empirical rule in detail, a standardized score called *z-value* must be introduced.

E) Comparing Variability With the Coefficient of Variation (CV)

The coefficient of variation, CV, is a standardized measure of dispersion in a distribution. It provides an index that measures the ratio of the standard deviation to the mean. The higher the CV, the greater the dispersion from the mean. It is used to compare the distribution of values for categories that have measurements are not directly comparable. Because it is a ratio, the CV can only be used with ratio data.

The CV can be reported as a decimal, percentage, or index.

The formula for the CV as an index:

Equation 11: Coefficient of Variation Equation

$$CV = \frac{\sigma}{\mu} * 100 \text{ or } \frac{s}{\bar{X}} * 100$$

Where: μ: Population mean
Σ: Population standard deviation
$\bar{X}$: Sample mean
s: Sample standard deviation

Some statisticians use a slightly different formula, and the results are reported as a percentage. Here is the alternative formula for reporting CY as a decimal:

Equation 12: Alternative Formula for the Coefficient of Variation

$$CV = \frac{\sigma}{\mu} \text{ or } \frac{s}{\bar{X}}$$

The coefficient of variation can be calculated in three simple steps:

Step 1: Find the mean and standard deviation.

Step 2: Divide the standard deviation by the mean.

Step 3: Multiply the result of step 1 by 100 (Optional).

Table 14 shows the prices for a Big Mac in 20 countries and Monthly Mobile data usage. These data sets use very different measures: Dollars and data usage in gigabytes. The CV will help determine whether the variance for the two data sets have equal variation.

Table 14: Coefficient of Variation – Big Mac Prices in US Dollars and Monthly Mobile Data Usage

Country	Big Mac Price in US$ (2018)	Monthly Mobile Data Usage in GB
Australia	$4.35	2.8
Bahrain	$3.18	6.7
Canada	$5.08	1.3
Britain	$4.07	1.9
China	$3.05	1.5
Czech Republic	$3.81	1.1
Denmark	$4.60	5.3
Hong Kong	$2.55	1.5
Hungary	$3.03	1.9
India	$2.55	1.5
Japan	$3.60	3.2
Poland	$2.80	3.4
Romania	$2.29	1.3
Singapore	$4.28	1.7
South Korea	$4.02	5.1
Sweden	$5.84	4.8
Switzerland	$6.62	3.0
Taiwan	$2.24	10.7
Turkey	$2.00	2.2
United States	$5.58	3.2

Sample Std. Dev.	$1.29	2.4
Sample Mean	$3.78	3.2
CV	34.29%	73.78%

Source: Big Mac prices, *The Economist*, January 2019. Mobile data usage, *tefficient*, July 2018.

By comparing the coefficients of variation, it is clear that monthly mobile data usage is far more variable than the price of a Big Mac. In fact, mobile data usage is nearly 2.2 times more variable than the price of a Big Mac, found by (73.78/34.29). **Please note:** These numbers for the sample standard deviation and sample mean have been rounded off to the hundredths column. This results is slightly different CVs than those calculated by using a handheld calculator, ($1.29/$3.78)*100 = 34.13%. This conclusion, however, is only a hypothesis. In Module 15, we will review a hypothesis test that compares variances from two samples.

Microsoft Excel does not have a built-in coefficient of variation formula, but you can use Excel to calculate the standard deviation and mean and complete the formula. To find out how this is done, open the file 05_CV_BigMac_Data.xlsx.

Closely related to the coefficient of variation is the coefficient of dispersion, which is the ratio of the variance to the mean.

IV. Measures of Relative Position

In addition to measures of dispersion there are also measures of relative position, which describe the placement of a particular value in a set of observations. These measures include *percentiles*, *quartiles*, *deciles*, *quintiles*, and standardized scores like *z-values*, *t-values*, and *F-values*. Standardized scores will be covered in detail in later modules.

As a measure of relative position percentiles, quartiles, deciles, and quintiles deal with ranked data. These measures, as we shall see, are closely related to the median.

- Percentiles divide the ordered data into 100 parts.

- Quartiles divide the ordered data into four parts.
- Quintiles divide the ordered data into five parts.
- Deciles divide the ordered data into ten parts.

The two most commonly used measures of relative position are percentiles and quartiles, which are *relative* scores. This means that these measures refer to the position of a variable in relation to all other variables. **Please note:** These values need not be actual values in the data.

Percentiles or Centiles

To repeat, percentiles are ranks where the ordered data is divided into 100 parts.

If a datum were located at the 75th percentile (P_{75}), this would be the point where 75 percent of the values fall at or below this position and 25 percent of the values would be above it. P_{50} is the middle value of the data and is, therefore, the median. Every value in a distribution is at or below the 100th percentile.

Calculating Percentile by Hand

The formula for calculating the percentile of ordered data by hand is:

Equation 13: Percentile Formula

$$P_i = (n+1)\frac{P_i}{100}$$

Where: P_i: Percentile with i standing for any particular percentile
N: Number of observations in the data

The first step is to order the data. Here is an ordered array of 12 values:

Table 15: Ordered array of 12 values

1	5	8	15	16	20	22	23	25	30	39	44

What is the 60th percentile? The formula is shown in Equation 14.

Equation 14: 60th Percentile Calculation

$$P_{60} = (12 + 1)\frac{60}{100} = 13 * 0.6 = 7.8$$

The 60^{th} percentile is the 7.8^{th} variable in the ordered array of twelve variables. It is 80 percent of the way between the seventh variable, 22, and the eight variable, 23. The 60^{th} percentile, therefore, is 22.8. **Remember:** The percentile need not be an actual value in the array.

Calculating Percentiles Using Microsoft Excel

Using Excel to calculate percentiles has big advantages over doing it by hand. It is faster because you do not have to calculate the distance between two of the variables. In addition, the data need not be arranged in numerical order.

Excel actually has three percentile functions:

1. **PERCENTILE:** This is an older "compatibility function, which returns the percentile. The syntax for this function is =PERCENTILE(array,k), where array is the cell reference for the data and k is the desired percentile from 0.0 to 1.0. **Please note:** The percentile, "k", must be entered as a decimal. The PERCENTILE function has been replaced by two new functions:
2. **PERCENTILE.EXC:** This function returns the k^{th} percentile in a range of value, where k is in the range of 0 to 1, exclusive. The syntax for this function is =PERCENTILE.EXC(array,k), where array is the cell reference for the data and k is the desired percentile from 0.0 to 1.0.
3. **PERCENTILE.INC:** This function returns the k^{th} percentile in a range of value, where k is in the range of 0 to 1, inclusive. The syntax for this function is

=PERCENTILE.INC(array,k), where array is the cell reference for the data and k is the desired percentile from 0.0 to 1.0.

Here is the 60th percentile calculation using the PERCENTILE.EXC function. **Please note:** The data need not be sorted in ascending or descending order.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1														
2													P_{60}	**Excel Formula**
3	**1**	**5**	**8**	**15**	**16**	**20**	**22**	**23**	**25**	**30**	**39**	**44**	22.8	=PERCENTILE.EXC(A3:L3,0.6)
4														
5	**23**	**25**	**30**	**39**	**44**	**1**	**5**	**8**	**15**	**16**	**20**	**22**	22.8	=PERCENTILE.EXC(A5:L5,0.6)
6														

Figure 16: 60th Percentile Calculated Using Excel

The data for Figure 16 can be found in the workbook 05_Percentiles_Quartiles.xlsx in the worksheet labeled Percentiles. **Please note:** The fact that you scored a 95 on your statistics exam does not mean your score is in the 95th percentile. Your percentile rank is based on the position of your score relative to all other test scores.

Quartiles

Quartiles divide the ordered data into four parts and are denoted by Q_1, Q_2, and Q_3. Quartiles and percentiles are closely related:

- Q_1 corresponds to P_{25}.
- Q_2 corresponds to P_{50} as well as the median.
- Q_3 corresponds to P_{75}.

Calculating Quartiles by Hand

There are two ways to find the quartiles *after the data have been arranged in numerical order*. The first way is to find three medians:

Step 1: Find the median for all the data. Q2 is the median of the entire data.

Step 2: Find the median of the data below Q2. This is Q1.

Step 3: Find the median of the data above Q2. This is Q3.

The second method is to employ this formula:

Equation 15: Quartile Formula

$$Q_i = Q_i\left(\frac{n+1}{4}\right)$$

Where: Q_i: Quartile, 1, 2, or 3, where $_i$ is the quartile
N: Number of observations in the data

Table 16: Ordered array of 12 values

1	5	8	15	16	20	22	23	25	30	39	44

Below are the quartiles for the ordered array shown in Table 16.

Here is the calculation for the Q1:

Equation 16: Q1 Calculation

$$Q_1 = Q_1\left(\frac{12+1}{4}\right) = 1\left(\frac{13}{4}\right) = 1 * 3.25 = 3.25$$

This means Q1 is the 3.25 variable, or 25 percent between 8 and 15. 25 percent of the area between 8 and 15 is 1.75, so Q1 is 9.75. found by 8 + 1.75.

Here is the calculation for the Q2:

Equation 17: Q2 Calculation

$$Q_2 = Q_2\left(\frac{12+1}{4}\right) = 2\left(\frac{13}{4}\right) = 2 * 3.25 = 6.50$$

This means Q2 is the 6.5 variable, or half way between 20 and 22, or 21.

Here is the calculation for the Q3:

Equation 18: Q3 Calculation

$$Q_3 = Q_3\left(\frac{12+1}{4}\right) = 3\left(\frac{13}{4}\right) = 3 * 3.25 = 9.75$$

This means Q3 is the 9.75 variable, or three-quarters of the way between 25 and 30 or 21. The distance between 25 and 30 is 5, 75 percent of 5 is 3.75. Q3 is 28.75, found by 25 + 3.75.

Calculating Quartiles by Microsoft Excel

As with percentiles, Excel accelerates the calculation process. Like the PERCENTILE functions, the data need not be ordered.

Excel has three quartile functions:

1. **QUARTILE:** This is an older "compatibility function. This formula returns the quartile. The syntax for this function is =QUARTILE(array,quart), where array is the cell reference for the data and can be one of five numbers: 0, 1, 2, 3, and 4. Zero returns the smallest value, 1 returns the first quartile, 2 returns the second quartile or the median, 3 returns the third quartile, and 4 returns the highest value. The QUARTILE function has been replaced by two new functions:
2. **QUARTILE.EXC:** This function returns the quartile for the data, based on percentile values from 0 to 1, exclusive. The syntax for this function is = QUARTILE.EXC(array,quart), where array is the cell reference for the data and quart can be one of five numbers: 0, 1, 2, 3, and 4. Zero returns the smallest value, 1 returns the first quartile, 2 returns the second quartile, 3 returns the third quartile, and 4 returns the highest value.
3. **QUARTILE.INC:** This function returns the quartile for the data, based on percentile values from 0 to 1, inclusive. The syntax for this function is = QUARTILE.EXC(array,quart), where array is the cell reference for the data and quart can be one of five numbers: 0, 1, 2, 3, and 4. Zero returns the smallest value, 1 returns the first quartile, 2 returns the second quartile, 3 returns the third quartile, and 4 returns the highest value.

Here are the quartiles found using QUARTILE.EXC.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1													**Q**	**Quartiles**	**Excel Formula**
2	**1**	**5**	**8**	**15**	**16**	**20**	**22**	**23**	**25**	**30**	**39**	**44**	1	9.75	=QUARTILE.EXC(A2:L2,1)
3													2	21.00	=QUARTILE.EXC(A2:L2,2)
4													3	28.75	=QUARTILE.EXC(A2:L2,3)

Figure 17: The QUARTILE.EXC Function

Please note: As with percentiles, the quartiles do not have to be actual values in the data.

The data for Figure 17 can be found in the workbook 05_Percentiles_Quartiles.xlsx on the worksheet labeled Quartiles.

Interquartile Range (IQR)

The interquartile range, also known as the IQR or middle-fifty, represents 25 percent of the data below the median and 25 percent above the median.

$$IQR = Q3 - Q1$$

Equation 19: IQR Formula

At the center of the IQR is the median, which is also Q2 and the 50^{th} percentile. The IQR is considered a better measure of variability in the data than the range because it is not affected by outliers.

The IQR is useful to identify what values are outliners. To see how this is done we turn to the *five number summary* and its graphic representation, called the *box-and-whisker chart*.

Five Number Summary and Box-and-Whisker Charts

The five number summary was invented by John W. Tukey as a method for describing the distribution of data. It describes the spread of the data around the median. The five numbers are:

1. The median.
2. The first quartile.

3. The third quartile
4. The smallest value in the sample.
5. The largest value in the sample.

This is not quite correct when there are outliers. Outliers are variables with very large or very small values that lie beyond where we expect to find data. There are, in fact, two kinds of outliers: Mild outliers and extreme outliers. Mild outliers are considered 1.5 times the interquartile range below the first quartile, or 1.5 times the interquartile range above the third quartile. Extreme outliers are considered 3.0 times the interquartile range below or above the IQR:

- Mild Outliers: Q1 – 1.5 * IQR or Q3 + 1.5 * IQR
- Extreme Low Outliers: Q1 – 3.0 * IQR or Q3 + 3.0 * IQR

Not all statisticians accept the use of the 1.5 and 3.0 multipliers. In 1987, two of Tukey colleagues, David C. Hoaglin and Boris Iglewicz, argued that the 1.5 multiplier leads to inaccurate results half the time while 3.0 is too stringent. They suggest using 2.2 instead.[7] We, however, will use the common 1.5 and 3.0 multipliers.

Box-and-Whisker Plots

Box-and-whisker plots, or box plots for short, are graphic representations of the five number summary. As we shall see, box and whisker plots are especially useful for finding outliers and comparing two or more data sets. At the center of these charts is a box that represents the interquartile range. The box has a line somewhere in the middle that represents the median. The bottom of the box is the start of the second quartile and on the top of the box is the end of the third quartile. The chart has two whiskers. The lower whisker covers the first quartile. It spans from the bottom of the second quartile to the

smallest value when there are *no outliers*. The upper whisker covers the fourth quartile. It spans from the top of the third quartile to the largest value when there are *no outliers*. See Figure 18 for a box-and-whisker plot with no outliers.

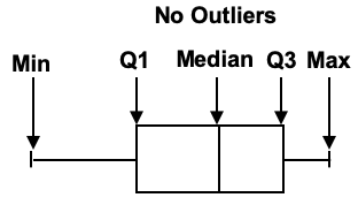

Figure 18: Generic Box-and-Whisker Plot Without Outliers

When there are outliers, whiskers extend to the largest or smallest variable that is *not* an outlier. The boundary between regular variables and outliers is marked with a dashed line called a *fence*.[8] Beyond this fence there is a dot for each outlier. These fences are sometimes called Tukey Fences after the creator of this graphic format. When there are extreme outliers an outer fence marks the border between outliers and extreme outliers. Beyond the outer fence dots or other symbols show the location of the extreme outliers.

All outliers should be investigated. An outlier may correctly represent the data or it may not.

Figure 19 shows what box-and-whisker plots look like when there are outliers.

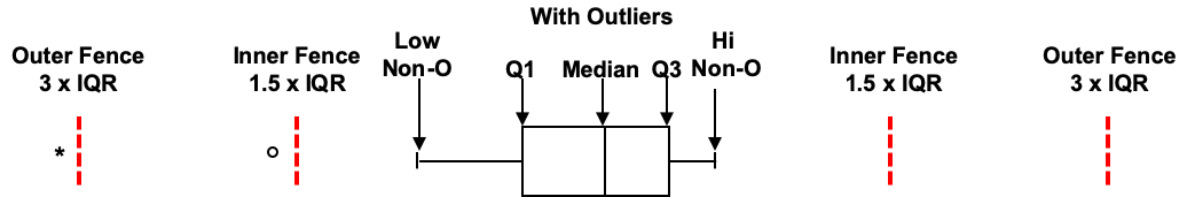

Figure 19: Generic Box-and-Whisker Plots With Outliers

Our Example:

For our five number summary and box-and-whisker plot we are going to compare the batting averages for two teams that played in Major League Baseball's American League's Division Series in 2018. The teams were the New York Yankees and the Boston Red Sox.[9]

This data can be found in the workbook titled 05_BoxAndWhisker.xlsx on the 2018ALDS worksheet.

Here are the batting averages for each player who had an “at bat” during this series. The data are sorted by teams.

	A	B	C	D
1	**NY Yankees**		**Boston Red Sox**	
2	**Player**	**BA**	**Player**	**BA**
3	Miguel Andujar	0.111	Andrew Benintendi	0.286
4	Brett Gardner	0.000	Mookie Betts	0.188
5	Didi Gregorius	0.214	Xander Bogaerts	0.294
6	Adeiny Hechavarria	0.000	Jackie Bradley Jr.	0.167
7	Aaron Hicks	0.200	Rafael Devers	0.286
8	Aaron Judge	0.375	Brock Holt	0.667
9	Andrew McCutchen	0.133	Ian Kinsler	0.308
10	Gary Sanchez	0.200	Sandy Leon	0.000
11	Gaiancarlo Stanton	0.200	J. D. Martinez	0.357
12	Gleyber Torres	0.308	Mitch Mooreland	0.333
13	Luke Volt	0.231	Eduard Nunez	0.182
14	Neil Walker	0.250	Steve Pearce	0.333
15			Blake Swibart	0.000
16			Christian Vasquez	0.333

Figure 20: 2018 ALDS: Player Batting Averages

Here is the five number summaries for both teams. The five numbers are in bold type:

Measure	NY Yankees	Boston Red Sox
Lowest	**0.000**	**0.000**
1st Q	**0.128**	**0.184**
Median	**0.200**	**0.290**
3rd Q	**0.236**	**0.333**
Highest	**0.375**	**0.667**
Mean	0.185	0.267
Count	12	14
IQR (3rd Q - 1st Q)	**0.108**	**0.150**
High Outlier	**0.000**	**0.557**
# of High Outlers	0	1

Figure 21: 2018 ALDS Five Number Summary

Included with the five numbers are the mean, the number of observations or count, the IQR, and the high outlier for the Boston Red Sox. Standard deviation was not included.

Here is the box-and-whisker plot.

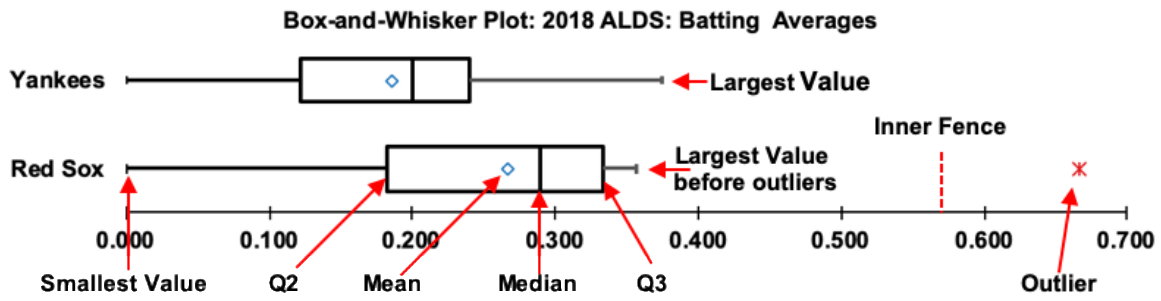

Figure 22: Box-and-Whisker Plot, 2018 ALDS

Low outlier fence .557, High outlier fence .783

As stated, all outliers should be investigated. One member of the Red Sox, Brock Holt, is an outlier with a .667 batting average. While this is a very high batting average, it is not unusual for a four-game series. It is two hits out of three at bats. An extreme outlier would be an average of .782 or higher. A .800 average is four hits out of five at bats. An extraordinary accomplishment, but certainly not impossible. Outliers like these can be checked. An outlier of over 1.000 would be impossible because it would represent more hits than at bats. If one of our players had such a high batting average, we should conclude that there is a serious problem with the data. If the source of the problem cannot be identified, we may decide to eliminate this variable.

This box-and-whisker plot quickly telegraphs that the Boston Red Sox had higher batting averages that the New York Yankees. We could perform hypothesis tests to confirm that the higher batting average for the Boston Red Sox was statistically significant. We need not do this. The sports press reported that the Red Sox's higher batting average had *practical significance*. The Red Sox beat the Yankees three games to one.

With Excel 2016, Microsoft introduced a box-and-whisker plot. Here is what this chart looks like using Excel's built-in chart:

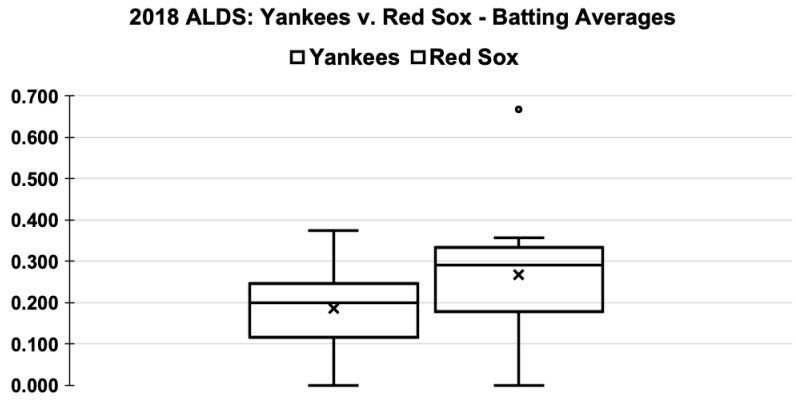

Figure 23: Result of Using Excel's Box-and-Whisker Plot

This chart is very easy to draw. All you have to do is highlight the data and go to "Insert," "Chart," and "Box and Whisker." There are, however, severe limitations with this chart. First, the chart can only have a vertical orientation. Many analysts want the option of a horizontal orientation like the chart shown in Figures 24 and 25. Second, the category labels are not shown on the X-Axis. To get around this problem, you must add a legend. Third, the program fails to draw a fence.

One way to overcome the limitations of Excel's new box-and-whisker function is to draw a box-and-whisker plot using a stacked bar or column chart. The rectangles for the first and fourth quartiles must be converted to "error bars," which are actually lines. This works when charting just one category. Unfortunately, you cannot create accurate whiskers when charting more than one category.

V. Estimating the Mean and Standard Deviation for Grouped Data

When data are placed into frequency distributions, they are "grouped." Grouping the data into separate classes, categories, buckets, or bins. makes the data easier to understand. But there is a cost. The datum-level detail necessary to calculate the mean and standard deviation with precision is lost. We can, however, estimate the mean and standard

deviation. Unfortunately, **Microsoft Excel does not have any built-in functions** to do this. But, you can enter the arithmetic functions in Excel to speed the estimation process.

Here is the frequency distribution constructed in Module 4. The frequency distribution summarizes 50 expense reports from clerical staff of the Dewey, Cheatem, and Howe law firm. When access to the raw data are not available, a question might arise: What are the mean and standard deviation? With the group data, a precise answer is not possible, but a reasonable estimate is feasible. Figure 24 shows the data for this problem.

	A	B
1	**Class**	**f**
2	**$110 < $120**	**4**
3	**$120 < $130**	**5**
4	**$130 < $140**	**9**
5	**$140 < $150**	**13**
6	**$140 < $150**	**9**
7	**$150 < $160**	**6**
8	**$160 < $170**	**4**
9	**Total**	**50**

Figure 24: Sample Frequency Distribution. Fictitious Data

These data for this frequency distribution can be found in 05_EstimatingMeanSD.xlsx.

A) Estimating the Mean

The mean for the data organized into a frequency distribution can be estimated using this formula:

$$\overline{X} = \frac{\Sigma fM}{n}$$

Equation 20: Formula for Estimating the Mean of Grouped Data

Where: $\bar{X}$: Estimated sample mean
Σ: Operation of addition
f: Class frequencies
M: Class mid-points
n: Number of observations

The class mid-point for the first class is found by adding the lower limit of the first class to the lower limit of the second class and dividing by two; $110 + $120 = $230/2 = $115. Once this mid-point is found, we add the class interval, $10, to this mid-point to find the class mid-point for the next class. This process is repeated until all the class mid-points are calculated. See Figure 25 for the frequency distribution with all the class mid-points.

	A	B	C
1	**Class**	**f**	**M**
2	**$110 < $120**	**4**	**115**
3	**$120 < $130**	**5**	**125**
4	**$130 < $140**	**9**	**135**
5	**$140 < $150**	**13**	**145**
6	**$140 < $150**	**9**	**155**
7	**$150 < $160**	**6**	**165**
8	**$160 < $170**	**4**	**175**
9	**Total**	**50**	

Figure 25: Frequency Distribution with Class Mid-Points. Fictitious Data

Class mid-points are used because the distribution of data within each class is unknown. Estimating that the average of the distribution will be at the mid-point is a way to get the most precise estimate of the mean without having access to the actual datum level distribution.

The next step is to multiply the class frequencies by the class mid-points, fM. Once the fM column has been added, divide these values by the number of observations, ΣfM/n. The best estimate of the mean is $145.40. See Figure 26.

	A	B	C	D
1	**Class**	**f**	**M**	**fM**
2	**$110 < $120**	**4**	**115**	**$460**
3	**$120 < $130**	**5**	**125**	**$625**
4	**$130 < $140**	**9**	**135**	**$1,215**
5	**$140 < $150**	**13**	**145**	**$1,885**
6	**$140 < $150**	**9**	**155**	**$1,395**
7	**$150 < $160**	**6**	**165**	**$990**
8	**$160 < $170**	**4**	**175**	**$700**
9	**Total**	**50**		**$7,270**
10			**n**	**50**
11			**Mean**	**$145.40**

Figure 26: Estimated Mean = $145.40
Fictitious Data

B) Estimating the Standard Deviation

You can estimate the standard deviation for the data organized into a frequency distribution using this formula:

$$s = \sqrt{\frac{\Sigma f(M - \bar{X})^2}{n - 1}}$$

Equation 21: Formula for Estimating the Standard Deviation of Grouped Data

Where: $\bar{X}$: Estimated sample mean
Σ: Operation of addition
f: Class frequency
M: Class mid-point
n: Number of observations

Here are the steps for estimating the standard deviation of data grouped in a frequency distribution.

Step 1: Estimate the mean. This has been done. The mean for our frequency distribution is $145.40.

Step 2: Subtract the Mean from the class mid-points. See Figure 27.

	A	B	C	D	E	F	G
1	**Class**	**f**	**M**	**fM**	**(M - Mean)**	**(M - Mean)^2**	**f(M - Mean)^2**
2	**$110 < $120**	**4**	**115**	**$460**	**-30.40**		
3	**$120 < $130**	**5**	**125**	**$625**	**-20.40**		
4	**$130 < $140**	**9**	**135**	**$1,215**	**-10.40**		
5	**$140 < $150**	**13**	**145**	**$1,885**	**-0.40**		
6	**$140 < $150**	**9**	**155**	**$1,395**	**9.60**		
7	**$150 < $160**	**6**	**165**	**$990**	**19.60**		
8	**$160 < $170**	**4**	**175**	**$700**	**29.60**		
9	**Total**	**50**		**$7,270**			
10			**n**	**50**			
11			**Mean**	**$145.40**			
12							

Figure 27: Subtract the Mean from the Class Mid-points
Fictitious Data

Step 3: Square the Deviations from the Mean, $(X - Mean)^2$. See Figure 28.

	A	B	C	D	E	F	G
1	**Class**	**f**	**M**	**fM**	**(M - Mean)**	**(M - Mean)^2**	**f(M - Mean)^2**
2	**$110 < $120**	**4**	**115**	**$460**	**-30.40**	**924.16**	
3	**$120 < $130**	**5**	**125**	**$625**	**-20.40**	**416.16**	
4	**$130 < $140**	**9**	**135**	**$1,215**	**-10.40**	**108.16**	
5	**$140 < $150**	**13**	**145**	**$1,885**	**-0.40**	**0.16**	
6	**$140 < $150**	**9**	**155**	**$1,395**	**9.60**	**92.16**	
7	**$150 < $160**	**6**	**165**	**$990**	**19.60**	**384.16**	
8	**$160 < $170**	**4**	**175**	**$700**	**29.60**	**876.16**	
9	**Total**	**50**		**$7,270**			
10			**n**	**50**			
11			**Mean**	**$145.40**			
12							

Figure 28: Square the Deviations from the Mean. Fictitious Data

Step 4: Multiply the Squared Deviations the Mean by the frequencies. See Figure 29.

	A	B	C	D	E	F	G
1	**Class**	**f**	**M**	**fM**	**(M - Mean)**	**$(M - Mean)^2$**	**$f(M - Mean)^2$**
2	**$110 < $120**	**4**	**115**	**$460**	**-30.40**	**924.16**	**3,696.64**
3	**$120 < $130**	**5**	**125**	**$625**	**-20.40**	**416.16**	**2,080.80**
4	**$130 < $140**	**9**	**135**	**$1,215**	**-10.40**	**108.16**	**973.44**
5	**$140 < $150**	**13**	**145**	**$1,885**	**-0.40**	**0.16**	**2.08**
6	**$140 < $150**	**9**	**155**	**$1,395**	**9.60**	**92.16**	**829.44**
7	**$150 < $160**	**6**	**165**	**$990**	**19.60**	**384.16**	**2,304.96**
8	**$160 < $170**	**4**	**175**	**$700**	**29.60**	**876.16**	**3,504.64**
9	**Total**	**50**		**$7,270**			
10			**n**	**50**			
11			**Mean**	**$145.40**			
12							

Figure 29: Multiply the Square Deviation from the Mean by the Frequencies. Fictitious Data

Step 5: Sum f(M – Mean)^2 and divide by n – 1 for the estimate of variance. Variance is estimated at 273.32. See Figure 30.

	A	B	C	D	E	F	G
1	**Class**	**f**	**M**	**fM**	**(M - Mean)**	**$(M - Mean)^2$**	**$f(M - Mean)^2$**
2	**$110 < $120**	**4**	**115**	**$460**	**-30.40**	**924.16**	**3,696.64**
3	**$120 < $130**	**5**	**125**	**$625**	**-20.40**	**416.16**	**2,080.80**
4	**$130 < $140**	**9**	**135**	**$1,215**	**-10.40**	**108.16**	**973.44**
5	**$140 < $150**	**13**	**145**	**$1,885**	**-0.40**	**0.16**	**2.08**
6	**$140 < $150**	**9**	**155**	**$1,395**	**9.60**	**92.16**	**829.44**
7	**$150 < $160**	**6**	**165**	**$990**	**19.60**	**384.16**	**2,304.96**
8	**$160 < $170**	**4**	**175**	**$700**	**29.60**	**876.16**	**3,504.64**
9	**Total**	**50**		**$7,270**		**Σ**	**13,392.00**
10			**n**	**50**		**n - 1**	**49**
11			**Mean**	**$145.40**		**s^2**	**273.31**
12							

Figure 30: Sum f(M - Mean)2 and divide by n – 1 to estimate variance. Fictitious Data

Step 6: Take the Square Root of the Estimated Variance to Calculate Standard Deviation. The estimate standard deviation is $16.53. See Figure 31.

	A	B	C	D	E	F	G
1	**Class**	**f**	**M**	**fM**	**(M - Mean)**	**$(M - Mean)^2$**	**$f(M - Mean)^2$**
2	**$110 < $120**	**4**	**115**	**$460**	**-30.40**	**924.16**	**3,696.64**
3	**$120 < $130**	**5**	**125**	**$625**	**-20.40**	**416.16**	**2,080.80**
4	**$130 < $140**	**9**	**135**	**$1,215**	**-10.40**	**108.16**	**973.44**
5	**$140 < $150**	**13**	**145**	**$1,885**	**-0.40**	**0.16**	**2.08**
6	**$140 < $150**	**9**	**155**	**$1,395**	**9.60**	**92.16**	**829.44**
7	**$150 < $160**	**6**	**165**	**$990**	**19.60**	**384.16**	**2,304.96**
8	**$160 < $170**	**4**	**175**	**$700**	**29.60**	**876.16**	**3,504.64**
9	**Total**	**50**		**$7,270**		**Σ**	**13,392.00**
10			**n**	**50**		**n - 1**	**49**
11			**Mean**	**$145.40**		**s^2**	**273.31**
12						**s**	**$16.53**

Figure 31: Estimated Standard Deviation for Grouped Data. Fictitious Data

VI. Descriptive Statistics with Excel's Descriptive Statistics Plug-In

Microsoft Excel's Analysis ToolPak enables users to run a variety of statistical analysis very quickly. One of the tools included in this plug-in is Descriptive Statistics. The descriptive statistics tool will quickly calculate 13 measures for multiple series of data:

1. Mean.
2. Standard Error (This very important concept will be presented in detail in later modules).
3. Median.
4. Mode (Unfortunately Excel does not record multiple modes).
5. Sample Standard Deviation.
6. Sample Variance.
7. Kurtosis.
8. Skewness. Note: This is calculated using the SKEW function.
9. Range.
10. Minimum.
11. Maximum.

12. Sum.

13. Count.

In addition, this tool can calculate confidence intervals for the mean and find the k^{th} smallest and largest values. We will review confidence intervals in detail in Module 11: Confidence Intervals and Estimating Parameters.

To launch this tool when using the Windows version of Excel, click on Data Analysis in the Data tab, and select Descriptive Statistics. To launch this tool in the Macintosh version of Excel, go to the Tools menu, select Data Analysis, and then select Descriptive Statistics.

Our example uses the batting averages for the 2018 American League Division Championship series. The data are shown in Figure 32.

	A	B
1	BAs 2018 ALDS	
2	**Yankees**	**Red Sox**
3	0.111	0.286
4	0.000	0.188
5	0.214	0.294
6	0.000	0.167
7	0.200	0.286
8	0.375	0.667
9	0.133	0.308
10	0.200	0.000
11	0.200	0.357
12	0.308	0.333
13	0.231	0.182
14	0.250	0.333
15		0.000
16		0.333

Figure 32: Batting Averages 2018 ALDC Series

When launching the descriptive statistics tool, the following dialog box appears (See Figure 33):

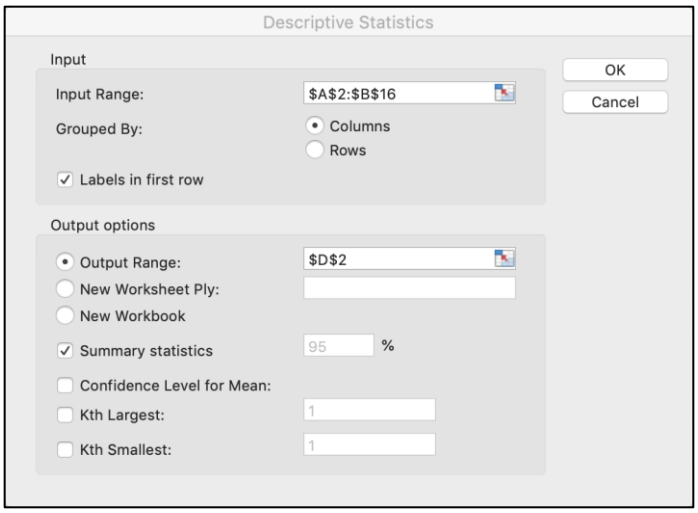

Figure 33: Descriptive Statistics Dialog Box

Under Input Range, enter the location of the data. By clicking on the Input Range box and then dragging the cursor through the appropriate cells. Because the data is organized on columns, click on the Grouped By Columns option. Select the output range, D2. Check the Summary statistics box. The other optional analyses are not checked.

In a few seconds, Excel calculates 13 measures for both teams. The output is shown in Figure 37. **Please note:** Excel does not format the numbers it reports. The numbers shown in columns D and F were manually formatted. This takes less than a minute. In cells H4:H16, I have added the measures for the Boston Red Sox calculated using Excel's built-in functions. The results are in n Column H. Column I shows the functions used. This data is available in the workbook titled 05_DescriptiveStat-TooPak.xlsx.

	A	B	C	D	E	F	G	H	I
1	**BAs 2018 ALDS**								
2	**Yankees**	**Red Sox**		*Yankees*		*Red Sox*			*Syntax*
3	0.111	0.286							
4	0.000	0.188		Mean	0.185	Mean	0.267	0.267	=AVERAGE(B3:B16)
5	0.214	0.294		Standard Error	0.032	Standard Error	0.044	0.044	=STDEV.S(B3:B16)/(SQRT(COUNT(B3:B16)))
6	0.000	0.167		Median	0.200	Median	0.290	0.290	=MEDIAN(B3:B16)
7	0.200	0.286		Mode	0.200	Mode	0.333	0.333	=MODE.SNGL(B3:B16)
8	0.375	0.667		Standard Deviation	0.111	Standard Deviation	0.164	0.164	=STDEV.S(B3:B16)
9	0.133	0.308		Sample Variance	0.012	Sample Variance	0.027	0.027	=VAR.S(B3:B16)
10	0.200	0.000		Kurtosis	0.043	Kurtosis	2.244	2.244	=KURT(B3:B16)
11	0.200	0.357		Skewness	-0.330	Skewness	0.524	0.524	=SKEW(B3:B16)
12	0.308	0.333		Range	0.375	Range	0.667	0.667	=MAX(B3:B16)-MIN(B3:B16)
13	0.231	0.182		Minimum	0.000	Minimum	0.000	0.000	=MIN(B3:B16)
14	0.250	0.333		Maximum	0.375	Maximum	0.667	0.667	=MAX(B3:B16)
15		0.000		Sum	2.222	Sum	3.734	3.734	=SUM(B3:B16)
16		0.333		Count	12	Count	14	14	=COUNT(B3:B16)

Figure 34: Descriptive Statistics Output

VII. Summary

We have completed our module on the common statistical measures. In Module 6 we will cover index numbers, which will complete our discussion of descriptive statistics. In Module 7 will be the basic concepts of probability. After that module we will move into inferential statistics.

VIII. Exercises

Data for these exercises can be found in 05_Exercises.xlsx.

Exercise 1:

	A	B	C	D	E	F	G	H
1	49	17	45	26	60		**Category**	**f**
2	19	99	98	40	51		**0 < 15**	**7**
3	3	41	81	68	45		**15< 30**	**12**
4	62	62	49	50	91		**30 < 45**	**8**
5	12	82	95	43	55		**45 < 60**	**14**
6	71	33	93	16	48		**60 < 75**	**11**
7	20	13	43	4	98		**75 < 90**	**9**
8	62	69	15	54	59		**90 < 105**	**9**
9	20	90	87	25	77		**Total**	**70**
10	53	77	28	63	45			
11	5	41	85	29	79			
12	20	71	12	69	54			
13	80	22	97	7	35			
14	91	35	83	55	64			

Question 1: Data

a) Estimate the mean using this frequency distribution, G1:H9.

b) Estimate the standard deviation using this frequency distribution, G1:H9

For the data shown in cells A1:E41, using Microsoft Excel calculate the:

c) Mode.

d) Media.

e) Mean.

f) Range.

g) MAD.

h) Population Variance.

i) Sample Variance.

j) Population Standard Deviation.

k) Sample Standard Deviation.

Exercise 2:

Using the data in C1:C8, calculate the following measures by hand:

a) Mode

b) Median

c) Mean

d) Range

e) MAD

	A	B	C
1			X
2		1	104
3		2	108
4		3	343
5		4	326
6		5	102
7		6	235
8		7	274

Question 2:Data

f) Population Variance
g) Sample Variance

h) Population Standard Deviation

i) Sample Standard Deviation

Check your answers using Microsoft Excel.

Exercise 3:

Using the data in C1:C9, calculate the following measures by hand:

a) Mode

b) Median

c) Mean

d) Range

e) MAD

f) Population Variance

g) Sample Variance

h) Population Standard Deviation

i) Sample Standard Deviation

Check your answers using Microsoft Excel.

	A	B	C
1			X
2		1	49
3		2	40
4		3	61
5		4	27
6		5	19
7		6	54
8		7	69
9		8	46

Question 3: Data

Exercise 4:

Based on your answers to Question 1, calculate Pearson's coefficient of skewness for this data.

Table 17: Question 4 Data

49	17	45	26	60
19	99	98	40	51
3	41	81	68	45
62	62	49	50	91
12	82	95	43	55
71	33	93	16	48
20	13	43	4	98
62	69	15	54	59
20	90	87	25	77
53	77	28	63	45
5	41	85	29	79
20	71	12	69	54
80	22	97	7	35
91	35	83	55	64

Exercise 5: A student received the following grades in her English class. Based on the weights of each assignment, what is this student's final average for the semester?

Table 18: Question 5 Data

Assignment	% of Final Grade	Grade
Reflective Essay 1	5.0%	74
Reflective Essay 2	5.0%	80
Term Paper 1	20.0%	82
Term Paper 2	20.0%	86
Mid-Term Exam	20.0%	75
Final Exam	20.0%	81
Class Participation	10.0%	82

Exercise 6: A first semester freshman received the following grades. What is her Grade Point Average for this semester?

Table 19: Question 6 Data

Course	Credits	Grade	Letter Grade	Grade Value
Pre-Calculus	3	1.70	A	4.00
Chemistry 1	4	3.30	A-	3.7
French 1	3	3.70	B+	3.30
English Composition 1	3	3.00	B	3.00
American History 1	3	3.00	B-	2.70
			C+	2.30
			C	2.00
			C-	1.70
			D+	1.30
			D	1.00
			D-	0.70
			F	0.00

Exercise 7: In July 2019, the Consumer Technology Association reported the following annual growth rates for the period between 2012 to 2019 for the U.S. Consumer Electronics category. Using Microsoft Excel, calculate the average annual growth rate for this eight-year period using the arithmetic and geometric means.

Table 20: Question 7 data

Year	Growth Rate
2012	4.6%
2013	2.2%
2014	3.8%
2015*	1.0%
2016**	1.5%
2017**	1.5%
2018**	6.0%
2019**	2.2%

Source: Consumer Technology Association, July 2019
**Estimated*
***Forecast*

Exercise 8: Look at the worksheet labeled Q8 in 05_Execises.xlsx. This worksheet shows weekly sales for each of the four new product designs in separate test markets. Using Microsoft Excel:

a) Construct a five-number summary

b) Calculate the means

c) Calculate the standard deviations

d) Calculate the trimeans

e) Draw box-and-whisker plots

f) Comment of your findings

Exercise 9: Figure 37 shows the number of fatal heart attacks in 19 selected European Union countries during 2014 along with the 2014 debt-to-GDP ratios for these countries in 2014.

- Compare the coefficients of variation and comment on your findings.

	A	B	C	D
1		Country	Fatal Heart Attacks 2014*	Debt to GDP Ratio**
2	1	Austria	14,521	225%
3	2	Belgium	7,792	327%
4	3	Czech Republic	26,171	128%
5	4	Denmark	3,943	302%
6	5	Finland	10,338	238%
7	6	France	33,513	280%
8	7	Germany	121,471	188%
9	8	Greece	12,200	317%
10	9	Hungary	32,138	225%
11	10	Ireland	4,283	390%
12	11	Italy	69,653	259%
13	12	Netherlands	8,956	325%
14	13	Poland	38,642	134%
15	14	Portugal	7,456	358%
16	15	Romania	50,667	104%
17	16	Slovakia	13,381	151%
18	17	Spain	32,564	313%
19	18	Sweden	12,617	304%
20	19	United Kingdom	69,325	252%

Figure 35: 19 Selected EU Countries - Fatal Heart Attacks and Debt to GDP Ratio

Exercise 10: Open the Q10 worksheet in 05_Exercises.xlsx. The worksheet shows two series: Group A and Group B. Run Microsoft Excel's Data Analysis Descriptive Statistics tool for both series.

a) Comment on your findings

b) Do your findings suggest that any further analysis is needed?

Exercise 11, The Knave and Fool Game: Washington Post, September 30, 2018. "Don't Be Fooled: Working Americans Are Worse Under Trump." By Robert J. Shapiro.

On September 30, 2018, Robert J. Shapiro, a senior fellow at Georgetown University's McDonough School of Business, wrote an op-ed for the *Washington Post*. Here is Mr. Shapiro's Op-Ed:

> ***The Washington Post***
> **Don't Be Fooled: Working Americans Are Worse Off Under Trump**
> Robert J. Shapiro, September 30, 2018
>
> *Robert J. Shapiro is the chairman of the advisory firm Sonecon and a senior fellow at Georgetown University's McDonough School of Business. He was President Bill Clinton's undersecretary of commerce for economic* ***affairs.***

Despite robust economic numbers during the Trump presidency, the American public has seemed curiously unmoved by such good news as the lowest U.S. unemployment level in nearly half a century. .. Its enthusiasm might have been dampened by this underappreciated economic reality: The typical working American's earnings, when properly measured, have declined during the Trump administration.

As any White House would, the president's economic team touts positive earnings data from the Bureau of Labor Statistics that suggest rising wages and salaries. But the figures are misleading. They focus not on how much an average working person earns but on the "average earnings" of all employed people. In times of rising inequality, employees at the top pull up "average" earnings. Shift to the bureau's earnings data for an average or "median: working person, and most of those claimed gains disappear. Another catch: The data used by the White House doesn't account for inflation. Adjust the median earnings data for inflation, and the illusion of progress evaporates.

Let's examine how much these technical sleights of hand distort what's happening to people's earnings. Using the White House's preferred data, *average* earnings rose from $894.06 in January 2017 to $937.02 in August 2018. That suggests impressive gains of $42.96 weekly over the 20-month period and $30.02 weekly over the past year. But what about median earnings rather than average earnings—that is, earnings of those in the middle of the distribution?

The Bureau of Labor Statistics has a different database for that view, and its quarter-by-quarter numbers show a very different picture. Median weekly earnings of all workers rose from $865 in the first quarter of 2017 to $876 in the quarter ending June 30, 2018. The typical working American's earnings increased $11 weekly over 18 months, barely more than one- quarter of the economic progress touted by the White House.

Even that modest gain is not very meaningful. The significance of what people earn lies in what they can do with their earnings, and inflation eats away at what any of us can purchase or save. As a result, serious earnings analysis is always framed in inflation-adjusted, or "real," terms. From January 2017 to June 2018, inflation totaled 3.77 percent, while the $11 increase in unadjusted weekly earnings over those 18 months represented gains of 1.27 percent.

The president does not control the economy, says columnist Catherine Rampell. No, really, he doesn't. {Gillian Brockell/The Washington Post}

To determine how much the real earnings of a typical working American fell during that period, simply adjust the $876 in median weekly earnings in the quarter ending June 30, 2018, for the 3.32 percent inflation that occurred in

the 18 months from the first quarter of 2017 to that date. The result: $876 in June 2018 had the same value as $848.20 in January 2017. In real terms, the weekly earnings of a typical working American fell $16.80, or 1.9 percent, during Donald Trump's first 18 months as president.

Another blow to the White House's preferred economic narrative: The current earnings decline is a new development. Using the same measure, real median weekly earnings increased substantially during Barack Obama's final 18 months as president.

Before adjusting for inflation, median weekly earnings increased during Obama's last 18 months from $803 in the third quarter of 2015 to $849jn the last quarter of 2016. . People's average weekly earnings thus increased $46, or 5.73 percent, before adjusting for inflation. Over the same months, cumulative inflation from July 2015 to December 2016 was 1.12 percent, so the real earnings of a typical working person clearly increased. By how much? Adjust the median weekly earnings in December 2016 of $849 for the 1.08 percent inflation over the preceding 18 months, which comes to $838.82. In real terms, the weekly earnings of a typical employed American increased $35.82, or 4.5 percent, over Obama's last 18 months in office, growing from $803 in the third quarter of 2015 to $838.82 in the fourth quarter of 2016.

In Ronald Reagan's succinct terms, average working Americans are worse off under the Trump presidency than they were under Obama's. Yes, low unemployment is something to applaud, but there might be a good reason that so many who have jobs aren't clapping.

Using your understanding of descriptive statistics, read this opinion piece. Answer the following questions.

In the opening paragraph, Mr. Shapiro writes, "The typical working American's earnings, when properly measured, have declined during the Trump administration."

A) How should the earnings of "typical working American's earnings" be measured?

B) Who, if anyone, is improperly measuring the earnings of typical working Americans? Are these people improperly measuring American's earnings because they are fools or knaves?

C) What data source does Shapiro and other analysts use? Is this a good source of data?

D) Who are the people Shapiro thinks are likely to be fooled?

E) Is Mr. Shapiro a knave, a fool, or an honest person making a reasonable argument?

F) Using the data that Shapiro and others use, how would you measure the typical American's earning under President Trump, President Obama, or any other president?

---

[1] John W. Tukey, *Exploratory Data Analysis*, (Reading, MA: Addison-Wesley, 1977), pp. 1-3.

[2] John W. Tukey, *Exploratory Data Analysis*, (Reading, MA: Addison-Wesley, 1977), p. 21)

[3] John W. Tukey, *Exploratory Data Analysis*, (Reading, MA: Addison-Wesley, 1977), p. 29.

[4] Baseball Hall of Fame, https://baseballhall.org/hall-of-famers/chadwick-henry.

[5] Peter H. Westfall, "Kurtosis as Peakedness, 1905-2014, RIP." *Journal of the American Statistician*, Vol. 68, No. 3, pp. 191-195. https://www.tandfonline.com/doi/abs/10.1080/00031305.2014.917055

[6] Ronald A. Fisher, "On the Mathematical Foundations of Theoretical Statistics." *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, Vol. 222, 1922, p. 311. *JSTOR*, www.jstor.org/stable/91208.

[7] David C. Hoaglin and Boris Iglewicz, "Fine-Tuning Some Resistant Rules for Outlier Labeling." *Journal of the American Statistical Association*, Vol. 82, No. 400. Pp. 1147-1149. December, 1987.

[8] John W. Tukey, *Exploratory Data Analysis*, (Reading, MA: Addison-Wesley, 1977), pp. 43-44.

[9] BaseballReference.com, https://www.baseball-reference.com/postseason/2018_ALDS1.shtml.